

DOCUMENT RESUME

ED 123 243

TM 005 294

AUTHOR Kalisch, Stanley James, Jr.
TITLE The Effects of Selected Variables and a Comparison of the Efficiency of Two Tailored Testing Procedures.
PUB DATE [Mar 75]
NOTE 67p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., April 1975); Not available in hard copy due to marginal legibility of original document; For a related document, see TM 005 232

EDRS PRICE MF-\$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS Ability; Branching; *Comparative Analysis; *Computer Oriented Programs; Decision Making; Efficiency; Instruction; *Models; Prediction; Probability; *Response Style (Tests); Simulation; Statistical Analysis; *Testing

IDENTIFIERS *Tailored Testing

ABSTRACT

Two tailored testing models, specifying procedures by which the correctness of examinees' responses to a fixed number of test items are predicted by presenting as few items as possible to the examinee, were compared for their efficiency. The models differ in that one requires reconsideration of each prediction whenever additional information is obtained, and the other does not. The results of computer simulations indicated that the testing procedures were more efficient when larger data bases were used. Losses incurred with small data bases were less than the losses that would be obtained if every item were presented to each examinee. (Author)

* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

4.02

THE EFFECTS OF SELECTED VARIABLES AND A
COMPARISON OF THE EFFICIENCY OF TWO
TAILORED TESTING PROCEDURES

Stanley J. Kalisch

Control Data Corporation
3255 Hennepin Avenue, South
Minneapolis, Minnesota 55408

A paper submitted at the annual meeting of
The American Educational Research Association
Washington, D. C.
March 31, 1975

This paper is a condensed form of a more detailed paper copyrighted under the title "The Comparison of Two Tailored Testing Models and the Effects of the Models' Variables on Actual Loss". Requests for copies of the more detailed paper should be sent to the author at the address given above.

© 1974 by Stanley James Kalisch, Jr.
All rights reserved.

PERMISSION TO REPRODUCE THIS COPY-
RIGHTED MATERIAL HAS BEEN GRANTED BY

S. J. KALISCH
TO ERIC AND ORGANIZATIONS OPERATING
UNDER AGREEMENTS WITH THE NATIONAL IN-
STITUTE OF EDUCATION. FURTHER REPRO-
DUCTION OUTSIDE THE ERIC SYSTEM RE-
QUIRES PERMISSION OF THE COPYRIGHT
OWNER.

ED123243

BEST COPY AVAILABLE

TM005 294

THE EFFECTS OF SELECTED VARIABLES AND A
COMPARISON OF THE EFFICIENCY OF TWO
TAILORED TESTING PROCEDURES

Two tailored testing models, specifying procedures by which the correctness of examinees' responses to a fixed number of test items are predicted by presenting as few items as possible to the examinee, were compared for their efficiency. The models differ in that one requires reconsideration of each prediction whenever additional information is obtained, and the other does not. The results of computer simulations indicated that the testing procedures were more efficient when larger data bases were used. Losses incurred with small data bases were less than the losses that would be obtained if every item were presented to each examinee.

¹This study was conducted at Florida State University, Tallahassee, Florida.

As a means of applying tailored testing to instructional measurement, Kalisch (1974) proposed a model which may be applicable to proper placement of individuals in an instructional sequence and serve to evaluate an individual's performance after instruction. The proposed model may also reduce testing time in that it provides a procedure for presenting a minimum number of items of a test and predicting the correctness of the responses to the remaining items. Kalisch (1974) has contended that the model is not dependent upon an assumed or validated hierarchy of skills and that the necessary number of responses to items by Ss, prior to implementation of the model, would be sufficiently small to provide accurate predictions and a practical means of successfully employing computer-based tailored testing. The model employs item conditional difficulties, beta distributions, and an expected loss function.

The four purposes of the study were:

1. To compare two versions of a tailored testing model similar to one suggested by Kalisch (1974).
2. To identify levels of the variables within the two versions, which produce an efficient tailored testing procedure.
3. To compare, within each version, the results obtained when employing relatively small data bases with those obtained using large data bases.
4. To determine, within each version, whether any of the tailored testing procedures are more efficient than obtaining the responses to all the items from each examinee.

THE TAILORED TESTING MODELS

Kalisch's (1974) model makes the following assumptions:

1. The probability of the correctness of an examinee's responses to items (item conditional difficulties) may be estimated by analyzing the response patterns of individuals who have previously answered the items and who have displayed the same item response pattern as the examinee.

2. The distribution of sample conditional difficulties of a specified item is approximated by a corresponding beta distribution,

The model proposed by Kalisch (1974) involved numerous variables which would necessarily be set equal to constants chosen by the examiners prior to implementation of the tailored testing procedure. Generally, the fewer the decisions to be made by the examiners the greater the ease in implementation. Hence, a revised version of Kalisch's earlier model is suggested. The revised model eliminates the need for some of the information previously required. The revised model will henceforth be referred to as Model 1.

In both the earlier model and Model 1, the decisions concerning prediction of the correctness of the examinee's item responses have not been reconsidered, even though additional information is obtained whenever the examinee answers an additional item. An adaptation of Model 1, hereafter referred to as Model 2, is identical to Model 1 with the exception that as additional information is obtained, previous decisions are reconsidered.

Conditional Difficulty of an Item

The models use a branching strategy and a decision making procedure that require determination of the probability that an examinee will answer a specified item correctly. The item difficulty provides such a probability value, but fails to make use of the information concerning the correctness of the examinee's responses to the previously presented items. The conditional difficulty of an item makes use of such information.

The conditional difficulty of an item for an individual is the probability of the individual's answering the k-th item correctly, given information concerning the correctness of the individual's responses to the preceding k-1 items and given prior subject-item response data. Hence, a conditional difficulty is a conditional probability.

Item Conditional Difficulty and the Beta Distribution

The tailored testing models being presented use population item conditional difficulties. Since the values obtained for the conditional difficulty of an item may vary according to the samples of examinees selected, the model considers the sampling variance of the conditional difficulty of an item. For any item with a sample conditional difficulty value between zero and one, there exists a beta distribution which has the following properties:

1. The expected value of the distribution equals the sample item conditional difficulty.
2. Its variance is nearly equal to the sampling variance of the conditional difficulty of the item.

3. Its domain is the closed interval from zero to one. It has been assumed that there exists a beta distribution which approximates the distribution of item conditional difficulty values obtained from infinitely many samples of examinees.

The models investigated use loss functions weighted by the probabilities of the population conditional difficulties of items being at each value in the closed interval $[0,1]$. Such probabilities may be determined by using a beta distribution. Since the area of the region between the horizontal axis and a beta distribution curve is one unit, the probability that the conditional difficulty of an item is in the interval $[a,b]$ is calculated from the formula

$$\begin{aligned} P(a \leq x \leq b) &= \int_a^b f(x) dx \\ &= \int_a^b \frac{(N-1)!}{(r-1)!(N-r-1)!} x^{r-1} (1-x)^{N-r-1} dx. \end{aligned} \quad [1]$$

Loss Functions

The tailored testing procedures being investigated require the prediction of the correctness of an examinee's responses to items, based upon his responses to previously presented items. If the population conditional difficulty of an item for an individual were equal to one, the decision maker should be willing to assume that the examinee would answer the item correctly if it were presented. A conditional difficulty equal to zero would warrant a decision to

assume the examinee's response to be incorrect. For population conditional difficulties between zero and one, the decision to predict the correctness of the examinee's response becomes a more difficult task.

Table 1 depicts a loss matrix applicable to the tailored testing models used. The dimension, information categories (Cronbach & Gleser, 1965), specifies two cases involving the population conditional difficulty p of an item: $p=1$ and $p=0$. The second dimension, the action, is separated into three levels: (1) predict the examinee's response to be correct, (2) predict the examinee's response to be wrong, and (3) obtain the examinee's response to the item.

TABLE 1

Matrix of Loss Values for Items with
Conditional Difficulty Values of 1 or 0

Action	Information Categories	
	$p = 1$	$p = 0$
Predict examinee's response to be correct	L_1	L_2
Predict examinee's response to be wrong	L_3	L_4
Obtain examinee's response to the item	L_5	L_6

The six values (represented by L_i 's, $i=1,6$) in the matrix are the expected losses for a specified action, given a particular information category (Hays & Winkler, 1970).

A Linear Loss Assumption

The tailored testing models involved in this study require the use of loss functions under conditions of uncertainty in decision making situations. An assumption in the models is that if the population conditional difficulty of an item is one, the examinee will answer the item correctly. Similarly, if $p=0$, it is assumed that the examinee will answer the item incorrectly. For the values of p between zero and one, the models assume linear loss functions, one within each of the two actions involving prediction.

A Non-Linear Loss Assumption

Although a linear loss assumption may be appropriate for the decisions involving prediction of a response, the same assumption does not seem appropriate for determining the loss associated with obtaining the S 's response. If the conditional difficulty of an item is either one or zero, the expected loss for obtaining a response would be expected to be higher than the losses associated with a prediction of the response. For items with conditional difficulties closer to the value of 0.5, it would be expected that the loss for obtaining a response would be less. A loss function appropriate for the action, obtain the student's response, appears to be a function

that is decreasing in the interval $[0, 0.5]$ and increasing in the interval $[0.5, 1]$. A function meeting these conditions is the quadratic function with an equation of the form $L(x) = k(x - 0.5)^2$, where $k > 0$.

Determining Expected Loss

As in Kalisch's (1974) model, the sample conditional difficulty of an item, p , is used as an estimate of the population parameter ρ . For each determined conditional difficulty between zero and one, there is a beta distribution whose expected value equals the estimated population conditional difficulty and whose variance closely approximates the sampling variance of the conditional difficulty. Within each action the loss value corresponding to each point in the interval $[0, 1]$ is weighted by the corresponding probability, determined from the beta distribution, and the weighted loss values are summed, to obtain the expected loss for the action. For all $N \geq 2$ and all r such that $0 < r < N$ there exists a beta distribution, β , as specified in Formula 7. Also, if for an action and for each value of x in the open interval $(0, 1)$ there exists a loss value $L(x)$, then the expected loss $E(A_i)$ for the action A_i , is given by the formula

$$E(A_i) = \int_0^1 L(x)\beta(x)dx. \quad [2]$$

where

$$\beta(x) = \begin{cases} \frac{(N-1)!}{(r-1)!(N-r-1)!} x^{r-1} (1-x)^{N-r-1} & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if } x < 0 \text{ or } x > 1 \end{cases} \quad [3]$$

Let L be a linear loss function defined by the equation

$$L(x) = L_i + (L_j - L_i)x, \quad [4]$$

where L_i and L_j are the loss values for an action for the conditions $p=0$ and $p=1$, respectively. For example, in Table 1, for the action of predicting an examinee's response to be wrong, L_i equals L_4 and L_j equals L_3 . Substituting Formulas 3 and 4 for $\beta(x)$ and $L(x)$, respectively, in Formula 2 and simplifying the formula produces the following result:

$$EL(A_i) = \frac{N-r}{N} L_i + \frac{r}{N} L_j. \quad [5]$$

For the decision, to obtain the examinee's response, the loss function employed is the quadratic relation of the form

$$L(x) = k(x - 0.5)^2, \text{ where } k > 0. \quad [6]$$

Let L_i and L_j be the loss values for the action for the conditions $p=0$ and $p=1$, respectively. Assume also that $L_i = L_j$. Hence, Formula 6 becomes

$$L(x) = 4L_i(x - 0.5)^2. \quad [7]$$

Substituting Formulas 3 and 7 for $\beta(x)$ and $L(x)$, respectively, in Formula 2 and simplifying the result produce the formula

$$EL(A_i) = L_i \left[4 \left(\frac{r}{N} \right) \left(\frac{r-N}{N+1} + 1 \right) \right]. \quad [8]$$

Selecting Loss Values for the Decision Loss Matrix

Kalisch's (1974) model required the decision maker to preset the critical conditional difficulty values V_1 and V_2 , which specified three possible conditions: $p < V_1$, $V_1 \leq p \leq V_2$, and $p > V_2$. These three conditions provided the decision maker with three discrete information categories for the decision loss matrix. The linear loss assumptions for the actions involving prediction and the quadratic loss assumption for the action, to obtain a response, have eliminated the need for the decision maker to supply both the critical conditional difficulty values and the loss values for the decision loss matrix. The loss values may be expressed as a function of the critical conditional difficulty values.

Consider the expected loss for predicting the correctness of a response as determined from Formula 5. Since the conditional difficulty of an item is $p = r/N$, Formula 5 may be written as

$$EL(A_i) = (1-p)L_i + pL_j. \quad [9]$$

Since L_i and L_j are constants, for any given value of p , the expected loss will be the same regardless of the actual values of r and N . Such constancy in expected loss values for given values of p is not the case for the expected loss for obtaining a response. The formula for expected loss for obtaining a response, as specified in formula 8, may be expressed in an alternate form,

$$EL(A_i) = L_i \left[4 \frac{p(p-1)}{1 + \frac{1}{N}} + 1 \right]. \quad [10]$$

The expected loss for obtaining a response is dependent not only upon the value of p but also upon the sample size N . For conditional difficulty values other than zero and one, the expected loss for obtaining a response decreases for a given conditional difficulty as N increases. As N increases and the conditional difficulty approaches 0.5, the value of the factor approaches zero. Intuitively, this result seems quite appropriate for predictive purposes since for a given conditional difficulty, the expected loss decreases as the size of the data base increases. Loss for this decision is minimized when the data base consists of responses from the entire population and the conditional difficulty equals 0.5.

In order to determine the loss values for the decision loss matrix, the following additional assumptions have been made:

1. The loss value for predicting a response to be correct when $p = 1$ equals zero.

2. The loss value for predicting a response to be wrong when $p = 0$ equals zero.

3. The loss value for predicting a response to be wrong when $p = 1$ equals the loss value for predicting a response to be correct when $p = 0$.

Figure 1 depicts the corresponding loss functions with respect to these additional assumptions. Using the notation L_p for the critical loss value for prediction, the notation L_o for the critical loss value for obtaining a response, and the three additional assumptions, Formulas 9 and 10 for expected loss may be written in the following forms:

1. For predicting a response to be correct,

$$EL(PC) = (1-p)L_p \quad [11]$$

2. For predicting a response to be wrong,

$$EL(PW) = pL_p \quad [12]$$

3. For obtaining a response,

$$EL(OR) = L_o \left[\frac{4p(p-1)}{1 + \frac{1}{N}} + 1 \right]. \quad [13]$$

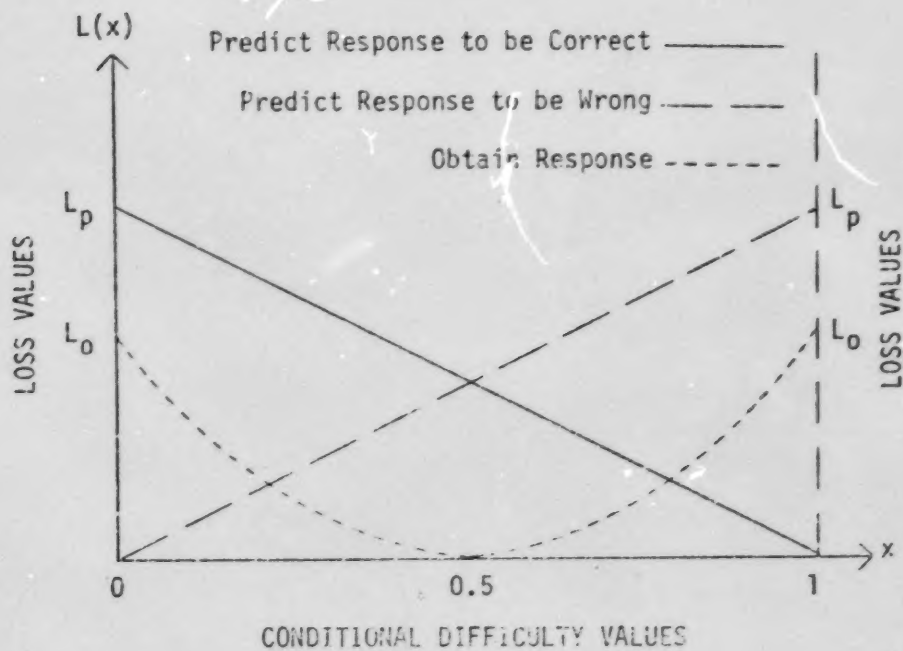


Fig. 1. The three loss functions used in the study.

One must now determine the values of L_p and L_o . Since the data base will not have responses from the entire population, it is necessary to specify the values of p such that $EL(PC) < EL(OR)$ and $EL(PW) < EL(OR)$. In other words, one must determine critical conditional difficulties such that the expected loss for predicting the correctness or incorrectness of a response is less than the expected loss for obtaining a response. The theory underlying the tailored testing models does not provide a means of determining the critical conditional difficulty values. Hence, the decision maker must supply these values. Under the present assumptions, the decision maker need only select one critical conditional difficulty value, V . The other critical value equals $1 - V$.

Suppose that the decision maker decides to select $V = 0.8$ in relation to a data base of size $N = 100$. For values of $p > V$, the decision maker desires the expected loss for predicting a response to be correct to be less than the expected loss for obtaining a response. Hence, from Formulas 11 and 13, $EL(PC) < EL(OR)$ implies

$$(1 - p) L_p < L_o \left[\frac{4p(p-1) + 1 + \frac{1}{N}}{1 + \frac{1}{N}} \right]. \quad [14]$$

Substituting 0.8 for p and 100 for N in Formula 14, the following relation is obtained

$$L_p < L_o (1.832). \quad [15]$$

Formula 15 specifies the conditions under which the loss for predicting a response as correct is less than the loss for obtaining the response. The values for L_p and L_o are relative. For convenience, let $L_o = 100$. Then $L_p = 183$ (to the nearest whole number). Hence, by specifying the critical conditional difficulty value of V and a base size N , values for L_p and L_o have been determined.

Model 1

The first tailored testing model that was investigated is specified by the following procedure:

1. A data base consisting of a subject by item matrix of

dichotomously scored items is obtained.

2. A critical conditional difficulty value and the corresponding data base size are selected. Using Formula 14, values for the critical loss, L_p and L_o , may be computed. The values of L_p and L_o are then used in the testing procedure to compute the expected loss for each decision, using Formulas 11, 12, and 13.

3. A minimum number of items to be presented to each S prior to any item response predictions for S is specified.

4. Responses from S to the minimum number of items, indicated in Step 3, are obtained. The branching strategy, used in the model, requires the difficulty values of the first items presented and the conditional difficulties of the subsequently presented items to be closest to 0.5.

5. For each unpresented item for which no prediction of the correctness of a response has been made, expected loss values for each possible action are determined using Formulas 11, 12, and 13. If the minimum of the three expected loss values indicates prediction to be the best decision, then the corresponding prediction is noted and the item is removed from further consideration. If the expected loss values for any two or more of the actions are equal to the computed minimum expected loss, a decision not to predict is made.

6. For each unpresented item for which no prediction has been made, that is, items for which the expected loss values were minimal for the action to obtain the examinee's response to the item, the conditional difficulty is determined. The branching strategy is

then employed. The item whose conditional difficulty is closest to 0.5 is then presented. After obtaining a response to the item, the procedure beginning with Step 4 is repeated.

7. As additional items are presented to S, the number of observations in the data base upon which conditional difficulties are computed is reduced. If the number of remaining observations in the data base is less than two, or if the denominator of the conditional difficulty ratios equals zero, the remaining items for which responses have neither been obtained nor predicted, are presented to S for his responses.

8. The model suggested by Kalisch (1974) noted that the maximum difference to be permitted between the actual sampling variance of the item conditional difficulty and the variance of the corresponding beta distribution could be specified. Kalisch's (1974) validation study did not employ a maximum difference. In the models investigated in this study, no maximum difference was specified. It was assumed that the corresponding beta distribution was a sufficiently close approximation of the assumed distribution of sample item conditional difficulty values for an examinee.

Model 2

In Model 1, in each case where a decision to predict the correctness of the examinee's response is made, no reconsideration of the prediction is made, even though additional responses have been obtained from the examinee. If the conditional difficulty of an item for an examinee equals either one or zero, additional information

will yield the same value for the conditional difficulty, r/N , provided there are sufficient observations to provide a non-zero denominator. For example, suppose that the conditional difficulty of a particular item equals $10/10$. Also suppose that after obtaining the response to another item with a conditional difficulty of 0.5 , it is found that the number of subjects in the data base, with an identical response pattern as the examinee's, equals five. Since all five subjects in the data base answered the former item correctly, the conditional difficulty equals $5/5$, or one. A similar situation exists for items with a conditional difficulty of zero. Item conditional difficulties between zero and one, may increase, decrease, or remain the same, as additional information is obtained. Model 2 is identical to Model 1 with the exception that the former involves recalculating the conditional difficulties and expected losses for items that have not attained values of zero or one, whenever the examinee supplies a response to an item.

The first four steps in the procedure specifying Model 1 also apply to Model 2. The remaining steps in Model 2 are as follows:

5. For each unrepresented item for which a previous conditional difficulty value, equaling neither zero nor one, exists, expected loss values for each possible action are determined, using Formula 10. If the minimum of the three expected loss values indicates prediction to be the best decision, and the conditional difficulty of the item equals zero or one, the corresponding prediction is noted and the item is removed from further consideration. If the expected loss values for any two or more of the actions are equal to

the computed minimum expected loss, a decision not to predict is made.

6. For each unpresented item for which the best decision was to obtain the examinee's response, the conditional difficulty is determined. Of these items, the one with a conditional difficulty closest to 0.5 is presented. After obtaining a response to the item, the procedure beginning with Step 5 is repeated.

7. If the number of remaining observations in the data base is less than two, or if the denominator of the conditional difficulty ratios equals zero, a terminating procedure is initiated. Predictions of the correctness of responses are noted for all items in which the last analysis of expected loss values indicated a prediction to be the best decision. The remaining items are presented to the examinee for his responses.

8. If during the tailored testing procedure, the best decision is a prediction of the correctness of the examinee's response to each of the remaining unpresented items, the predictions are noted and the testing procedure is terminated.

THE PROBLEM

As previously stated, the purpose of the proposed study was to compare the two tailored testing models, Model 1 and Model 2, and investigate the levels of the variables within the two versions, in terms of actual losses. It was also a purpose of the study to determine if the tailored testing procedures are efficient, that is, whether the total actual losses incurred by employing the models are less than the losses which would be incurred by presenting all the items to each examinee.

Specifically, the following questions concerning the tailored testing models, Model 1 and Model 2, were to be answered by the study:

1. Which of the two versions of the tailored testing models is more efficient?
2. How many items should be presented to the examinee before prediction of the correctness of a response is attempted?
3. What critical conditional difficulty values should be selected by the decision maker?
4. Are the tailored testing procedures using data bases consisting of fewer than 1000 sets of item responses as efficient as the testing procedures using data bases consisting of 2000 sets of

responses?

5. Are the tailored testing procedures using data bases consisting of fewer than 1000 sets of item responses more efficient than presenting all the items to an examinee?

METHODS AND PROCEDURES

Computer simulations employing existing data were used for the study. Computer programs were written by the investigator, using FORTRAN IV and COMPASS languages, and were executed on the Control Data Corporation (CDC) 6500 Computer at Florida State University.

The data used in the study were collected in 1973 through the Florida Statewide Assessment Program, which is an objective-referenced testing program (see Footnote 2). One area of testing was ninth grade mathematics, in which three different tests were administered. Each of approximately 15,000 Ss, who were selected by use of matrix sampling procedures, received one of the three tests.

Every item of each test was analyzed by the investigator for readability, correctness, and content validity in relation to the specified corresponding objectives. No two tests contained items corresponding to the same objective, and each objective had more than one item on a test. After removing from consideration ambiguous items and items lacking content validity, nine objectives and the corresponding 33 items of one of the test forms were selected because of the homogeneous content and the inter-relation of the skills required for mastery of the objectives. For the items

²Florida Statewide Assessment Program, 1973, State Department of Education, Tallahassee, Florida.

selected, 4913 Ss provided responses to at least one of the items. Some examinees apparently did not have a sufficient amount of time to complete answering all the items, as evidenced by the absence of responses to consecutive items at the end of the test. All sets of responses in which no response was obtained to two or more of the latter consecutive items were removed from consideration. Responses given by 4844 Ss remained. For the sets of responses remaining, any absence of a response to an item by an examinee was interpreted as an incorrect response.

Using program TESTAT (Veldman, 1967), it was determined that the mean number of correct responses to the 33 items by the 4844 Ss was 24.32; the standard deviation was 9.18; the reliability of the test scores, using KR-20, was .94; and the difficulties of the items ranged from .41 through .93.

The study consisted of two parts. At the beginning of each of the two parts of the experiment, each S's set of item responses was randomly assigned to one of two sets. Throughout each part of the study, Ss' data needed for a data base were selected from one set and the Ss' data to be used for the hypothetical Ss receiving the tailored tests, were selected from the other set.

Dependent Variables

To compare the results of each tailored testing procedure, total actual losses were computed. The sum of the actual losses for each item constitutes the total actual loss for the examinee. To determine loss per item, evaluation loss matrices were used. Three such matrices were employed. One matrix was specified by the investigator and the others were obtained independently from two individuals knowledgeable in the area of computer-based instruction and testing. The dependent variables were the total actual losses per examinee, in relation to each of the three evaluation matrices, which are exhibited in Tables 2, 3, and 4 and which are referred to as Evaluation Matrices A, B, and C, respectively.

Some of the questions that were to be answered by the study have been stated using the term "efficiency." Efficiency has been defined in terms of total actual loss. If procedure A yields lower total actual losses than procedure B, with actual losses computed on the basis of the same evaluation loss matrix as procedure A, then procedure A is said to be more efficient than procedure B.

TABLE 2

Evaluation Loss Matrix A

Action	Item Response, if Obtained, Would Be	
	Correct	Wrong
Predict examinee's response to be correct	1	20
Predict examinee's response to be wrong	15	1
Obtain examinee's response to the item	5	5

TABLE 3

Evaluation Loss Matrix B

Action	Item Response, if Obtained, Would Be	
	Correct	Wrong
Predict examinee's response to be correct	1	10
Predict examinee's response to be wrong	5	1
Obtain examinee's response to the item	20	15

TABLE 4
Evaluation Loss Matrix C

Action	Item Response, if Obtained, Would Be	
	Correct	Wrong
Predict examinee's response to be correct	0	3
Predict examinee's response to be wrong	2	0
Obtain examinee's response to the item	1	1

Independent Variables

The four independent variables considered were (1) the models, (2) the critical conditional difficulty values, (3) the number of items to be presented prior to prediction, and (4) the number of Ss for the data base.

The models. As previously mentioned, the tailored testing models, Model 1 and Model 2, were used.

The critical conditional difficulty values. With respect to the assumptions that have been made for this study, the specification of one critical conditional difficulty V , between zero and one, determines the value of the other critical conditional difficulty value, $1-V$. As previously discussed, the critical conditional

difficulty values are dependent upon the size of the data base upon which predictions are being made. In the study being reported, all specifications of critical conditional difficulties assumed the size of the data base to be equal to 10. The critical conditional difficulties selected for the investigation were .80, .85, .90, and .95. Using Formula 14 and letting L_0 , the critical loss value for obtaining a response, equal 100, a value for L_p , the critical loss value for predicting the correctness of a response, was determined for each of the selected critical conditional difficulty values.

With the critical loss values determined, a decision loss matrix such as in Table 1 may be created, by substituting zero for L_1 and L_4 , 100 for L_5 and L_6 , and the critical loss values (209, 358, 673, or 1655 corresponding to critical conditional difficulty values of .80, .85, .90, and .95, respectively) for L_2 and L_3 .

The levels of the critical conditional difficulties of .80, .85, .90, and .95, referred to as Level 1, Level 2, Level 3, and Level 4, respectively, were coded for computational purposes as 1, 2, 3, and 4.

Number of items to be presented prior to predictions. In both Model 1 and Model 2, the decision maker must specify how many items will be presented to the examinee prior to any attempt at prediction. The values of concern were limited to the interval [1,5].

Number of subjects for the data base. As previously discussed, the practicality of Model 1 and Model 2 may depend upon the number of S_s from whom responses to all the test items are necessary. The number of S_s to be considered for the data base was restricted to values equal to multiples of 100 between 100 and 2000, inclusive.

Part 1 of the Study

As a means of determining the effects of the versions of the models, the critical conditional difficulty levels, and the number of items presented prior to predictions, a three-factor partially nested univariate analysis of variance was employed for each of the three dependent variables--total actual loss per examinee, as determined from the three evaluation matrices. The mathematical model used was

$$Y_i = \mu + V_j + C_{k(j)} + I_m + VI_{jm} + CI_{km(j)} + \epsilon_{p(jkm)} \quad [16]$$

where Y_i = total actual loss per examinee ($i=1,3$)
 V_j = version of the model ($j=1,2$)
 $C_{k(j)}$ = critical conditional difficulty level ($k=1,4$)
 I_m = number of items presented prior to attempted predictions ($m=1,5$).

The factor, critical conditional difficulties, was nested within versions of the model because of the different roles played by the critical conditional difficulties in the two models. In Model 2, the prediction of the correctness of an item response is conditional, that is, the prediction is reconsidered each time another response is obtained from the examinee. Reconsideration of a prediction does not occur in Model 1.

For each of the 40 cells a data base consisting of the responses given by 2000 randomly selected Ss to the 33 items was formed. For each of the 40 cells a validation group consisting of 25 randomly selected Ss was also formed.

For each of the six cases determined by the version of the model and the evaluation matrices, the levels of the critical conditional difficulty and the number of items presented prior to prediction that produced the lowest point estimate of total actual loss were determined. The levels were determined by employing regression analysis. On the basis of the three analyses of variance performed, if significant main effects for the conditional difficulty level (C) and the number of item presented to attempted prediction (I) were indicated, but the CI interaction was non-significant, the following model was used:

$$Y = B_0 + B_1C + B_2I + \epsilon. \quad [17]$$

For each regression equation obtained, a lack of fit test and significance of coefficients test were performed ($\alpha = .05$). If

it was concluded that the model fit the data and each of the coefficients B_1 and B_2 were significantly different from zero, the model was assumed to adequately represent the data. Regression equations failing to fit the data or failing to have both coefficients significantly different from zero were judged as inadequate. For all cases in which the model given by Formula 17 was not applied and for all cases in which the same model was inadequate, the following model was used:

$$Y = B_0 + B_1 C' + B_2 I' + B_{11} C'^2 + B_{22} I'^2 + B_{12} C' I' + \epsilon, \quad [18]$$

where $C' = C^u$ and $I' = I^v$.

Rather than employing polynomials beyond the second order to provide a more adequate fit of the data to a model and to obtain, if possible, estimates of regression coefficients that were significantly different from zero, a reiterative procedure developed by Box and Tidwell (1962) was used. The procedure provides estimates of power or logarithmic transformations of the independent variables to reduce the residual sums of squares. The model used for this procedure is given by Formula 18. Program BMD03R (Dixon, 1971) was used to obtain the estimates of the coefficients and the residual for each observation. These values were used as input to programs written by the investigator to obtain estimates of the power or logarithmic transformations of the independent variables, which were then available as input into BMD03R for determination of another regression equation. The procedure was continued until convergence of the coefficients was indicated or until transformations for the independent variables produced excessively large numbers or numbers very equal to zero.

Using the regression equations based on the linear model, if adequate, and the equations obtained from the reiterative procedure, point estimates of the total actual loss at each of the 20 design points were determined. Within each of the six version by evaluation matrix categories, the levels of the independent variables yielding the minimum total actual loss were identified for use in the second part of the study.

Part 2 of the Study

With the levels of the variables set to values identified in part 1 as producing the most efficient procedures, it was to be determined if equal efficiency would be obtained for smaller data bases. The design used to answer this question was a repeated measures design. Sets of responses were randomly selected. These sets of responses served as the responses of hypothetical Ss comprising the examinee group. Two sets of responses were independently and randomly selected, one containing 2000 sets of responses and the other containing 100 sets. Using the prescribed levels of the variables, two independent simulations of the tailored testing procedure were conducted for the examinee group, using one data base in one simulation and the other data base in the other simulation. A one-tailed t-test for determination of a difference between two means for correlated samples was used. Let $\bar{L}_{2000}$ and $\bar{L}_{100}$ represent the total actual loss means for the simulations involving the data bases with 2000 and 100 sets of responses,

respectively. The null hypothesis was

$$H_0: \bar{L}_{100} - \bar{L}_{2000} = 0$$

with respect to the alternate hypothesis

$$H_1: \bar{L}_{100} - \bar{L}_{2000} > 0.$$

The investigator, applying decision theory to the rejection or non-rejection of H_0 (Hays & Winkler, 1970), was either to accept H_0 or to accept H_1 . In order to use the decision theory approach, careful consideration was given both to Type I and Type II errors.

In the context of the example being discussed, the effect size index, d , for t -tests of means is defined (Cohen, 1969) as

$$d = \frac{\bar{L}_{100} - \bar{L}_{2000}}{\sigma} \quad [19]$$

where σ is the common within-population standard deviation.

Selecting $\alpha = .05$, $\beta = .15$, and $d = .2$, the required number of paired observations may be determined from power tables (Cohen, 1969, pp. 28-29). Approximately 363 paired observations would have been necessary had the paired observations been uncorrelated. Since the paired observations were to be correlated, an adjusted effect size value, d_1 , is determined from the formula

$$d_1 = \frac{d}{\sqrt{1-r}}, \quad [20]$$

where r is the correlation between the paired observations. Preliminary simulations indicated that the value of r generally exceeded .7. Using the power tables, it was determined that if $\alpha = .05$, $\beta = .15$, $d = .2$, and the number of paired observations to be used

equalled 100, d_1 would equal .385 and r would approximately equal .68. Hence, in order to insure a Type II error of at most .15, it was required that the correlation between the 100 paired observations be greater than or equal to .68.

If H_0 was not rejected, it was to be assumed that the tailored testing models employing the data bases of different sizes were equally efficient. If H_0 was rejected in favor of H_1 , it was to be accepted that the procedure employing the smaller data base was less efficient than the procedure using the larger data base. In the case that H_0 was rejected, a simulation involving the 100 sets of responses comprising the examinee group and a randomly selected data base consisting of 200 sets of responses was to be conducted, and a comparison of total actual losses would be made for the 2000-data base and the 200-data base simulations. If the 200-data base failed to provide as efficient a procedure as the 2000-data base, a data base containing 1000 sets of responses was to be used. If a 1000-data base did not provide as efficient a procedure as the 2000-data base, no other data bases were to be considered for that version by evaluation category. If the 1000-data base provided as efficient a procedure as the 2000-data base, the investigation was to be continued. The smallest data base, which contained an integral multiple of 100 responses, between 200 and 1000, and which provided as efficient a procedure as the 2000-data base, was to be identified.

The final purpose of the study was to determine for each of the six versions of the model by evaluation matrix categories

whether a tailored testing procedure employing a data base of fewer than 1000 sets of responses would be more efficient than presenting all the items. In relation to Evaluation Matrices A and C, shown in Tables 2 and 4, the total actual losses for presenting 33 items are 165 and 33, respectively. For Evaluation Matrix B, shown in Table 3, the minimum total actual loss for presenting 33 items occurs if all the items are answered incorrectly. Hence, in relation to Evaluation Matrix B, the minimum total actual loss for presenting all the items is 495.

For each of the six cases, a t -test was conducted to determine the difference between the total actual loss for the tailored testing procedure and the actual loss for presenting all the items. Data bases of size 100 were initially considered. In the same way as for the previous set of t -tests, the smallest data bases of sizes between 100 and 1000, that were more efficient than presenting all the items were to be identified. Consider the following example. Let $\bar{L}_{100}$ equal the mean total actual loss incurred from a tailored testing procedure, using a base of size 100. Let L represent the total loss incurred by presenting all the items--165, 495, or 33, depending upon the evaluation matrix involved. The null hypothesis tested was

$$H_0: \bar{L}_{100} - L = 0$$

with respect to the alternate hypothesis

$$H_1: \bar{L}_{100} - L < 0.$$

In a practical situation, a computer-based testing program would probably not be initiated unless a moderately large difference between $\bar{L}_{100}$ and L existed. Hence, for the t -tests to be performed for this part of the study, an effect size of $d = .7$ was selected. With a selected value of $\alpha = .01$, the number of observations equal to 100, and $d = .7$, β is less than .005.

RESULTS

Part 1 of the Study

Descriptive statistics for the total actual loss for Model 1 and Model 2 are shown in Tables 5 and 6, respectively. For each of the 40 model by critical conditional difficulty level by items presented prior to attempting prediction categories, three total actual loss means were computed, according to the values in the three evaluation matrices. Hence, the statistics in any one row of Table 5 or Table 6 are based upon the results of the same simulation but reflect the differing evaluation criteria implied by the values in the evaluation matrices.

Descriptive statistics of the results of the first part of the study are shown in Tables 7 and 8. Included in Tables 7 and 8 are the means, standard deviations, and ranges for the number of items presented to each hypothetical examinee, the number of responses correctly predicted per examinee, and the number of responses incorrectly predicted per examinee, for each of the 40 simulations conducted according to the procedures specified by Model 1 and Model 2.

TABLE 5
Descriptive Statistics of Total Actual Loss for
Model 1 in Part 1 of Study

Minimum Number of Items to be Presented	Evaluation Loss Matrix					
	A		B		C	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Critical Conditional Difficulty--Level 1						
1	140.44	64.60	217.44	58.94	20.04	10.55
2	148.88	74.14	212.08	51.72	21.32	11.96
3	125.84	45.29	213.84	53.92	17.92	7.65
4	148.08	71.58	234.68	58.72	21.68	11.53
5	145.40	74.58	240.20	66.12	21.52	12.32
Critical Conditional Difficulty--Level 2						
1	117.48	40.51	224.48	61.08	17.00	7.21
2	139.68	72.56	253.08	59.08	21.00	11.72
3	92.68	30.33	205.48	67.13	12.84	5.82
4	145.76	64.19	263.76	69.69	22.08	10.70
5	122.08	57.94	233.48	81.19	17.92	10.29
Critical Conditional Difficulty--Level 3						
1	129.84	41.98	259.04	66.71	19.68	7.56
2	131.64	50.67	274.64	93.27	20.32	9.35
3	128.72	50.16	270.92	67.32	19.80	8.91
4	137.20	55.18	274.40	72.89	21.16	9.37
5	134.40	53.00	271.80	69.91	20.56	9.05
Critical Conditional Difficulty--Level 4						
	132.16	50.27	286.96	55.48	20.68	8.61
	125.04	60.64	282.44	55.80	19.40	9.81
	127.88	40.17	300.88	59.93	20.36	7.04
4	129.76	50.93	282.55	67.61	20.20	8.82
5	130.72	37.59	289.72	49.65	20.44	6.41

TABLE 6
Descriptive Statistics of Total Actual Loss for
Model 2 in Part 1 of Study

Minimum Number of Items to be Presented	Evaluation Loss Matrix					
	A		B		C	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Critical Conditional Difficulty--Level 1						
1	124.28	43.38	227.68	71.71	18.08	7.65
2	132.68	57.86	239.88	86.16	19.64	10.28
3	128.72	54.05	240.92	88.04	18.96	9.64
4	132.96	55.20	257.16	85.91	19.96	9.89
5	137.48	61.49	260.48	92.76	20.76	10.90
Critical Conditional Difficulty--Level 2						
1	125.68	44.64	268.68	90.66	19.20	8.38
2	126.72	53.49	249.92	76.42	18.92	9.29
3	129.24	48.63	255.24	82.26	19.40	8.81
4	129.92	49.22	255.72	77.67	19.56	8.81
5	120.72	40.02	276.32	83.97	18.72	7.56
Critical Conditional Difficulty--Level 3						
1	102.36	41.75	243.16	71.13	15.12	7.68
2	101.40	37.51	247.60	81.38	15.12	7.36
3	120.64	31.63	272.64	66.32	18.56	5.98
4	141.20	57.98	296.60	74.28	22.08	9.98
5	134.72	54.69	302.92	87.35	21.40	9.59
Critical Conditional Difficulty--Level 4						
1	136.16	43.76	315.36	62.98	21.96	7.59
2	130.08	45.32	308.68	41.55	20.76	7.44
3	138.32	42.60	306.52	66.60	21.88	7.36
4	128.88	47.51	295.28	61.67	20.40	8.33
5	137.08	41.59	316.68	65.23	22.12	7.23

TABLE 7

Descriptive Statistics of the Results for Model 1 in Part 1 of the Study

Number of Items Before Predic- tion	Category								
	Items Presented			Correct Predictions			Incorrect Predictions		
	Mean	S.D.	Range	Mean	S.D.	Range	Mean	S.D.	Range
Critical Conditional Difficulty--Level 1									
1	9.08	3.24	[4,15]	19.84	5.35	[10,28]	4.08	3.38	[0,14]
2	8.80	2.47	[4,14]	19.68	4.98	[7,29]	4.52	3.80	[0,17]
3	9.76	3.05	[4,16]	20.04	4.85	[9,29]	3.24	2.39	[0, 8]
4	10.16	3.38	[5,16]	18.68	5.26	[6,27]	4.20	3.77	[0,18]
5	10.76	3.67	[5,18]	18.40	5.91	[9,28]	3.88	3.76	[0,16]
Critical Conditional Difficulty--Level 2									
1	10.44	3.61	[5,18]	20.08	4.72	[12,28]	2.48	1.90	[0, 8]
2	11.64	3.51	[7,18]	18.08	5.17	[6,26]	3.32	3.76	[0,15]
3	9.52	4.20	[5,24]	22.28	4.69	[9,28]	1.24	1.36	[0, 5]
4	12.40	4.20	[5,20]	16.96	5.75	[9,28]	3.68	3.25	[0,14]
5	10.80	4.75	[5,22]	19.68	6.41	[9,29]	2.56	2.64	[0, 8]
Critical Conditional Difficulty--Level 3									
1	12.12	4.25	[5,21]	18.24	5.09	[8,27]	2.64	1.94	[0, 8]
2	13.68	5.88	[5,27]	16.64	7.06	[4,28]	2.68	2.34	[0, 8]
3	13.40	3.93	[6,22]	17.12	5.57	[6,27]	2.52	2.30	[0, 9]
4	13.48	4.54	[6,25]	16.60	5.69	[2,27]	2.96	2.93	[0,11]
5	12.80	4.20	[6,20]	17.40	5.42	[4,27]	2.84	2.63	[0,11]
Critical Conditional Difficulty--Level 4									
1	13.48	3.59	[10,23]	16.76	4.84	[6,23]	2.40	2.42	[0,10]
2	13.44	3.20	[9,19]	17.44	4.46	[7,23]	2.12	3.18	[0,16]
3	14.96	3.89	[10,23]	16.08	4.78	[7,23]	2.00	2.00	[0, 8]
4	13.72	4.83	[9,27]	16.96	5.64	[5,24]	2.36	2.53	[0,11]
5	13.80	3.16	[9,21]	16.92	3.87	[9,24]	2.32	1.83	[0, 6]

TABLE 8

Descriptive Statistics of the Results for Model 2 in Part 1 of the Study

Number of Items Before Predic- tion	Category								
	Items Presented			Correct Predictions			Incorrect Predictions		
	Mean	S.D.	Range	Mean	S.D.	Range	Mean	S.D.	Range
Critical Conditional Difficulty--Level 1									
1	10.64	4.94	[4,20]	19.48	5.71	[10,29]	2.88	2.29	[0, 7]
2	11.16	5.35	[4,19]	18.68	6.79	[6,29]	3.16	2.66	[0,11]
3	10.96	5.23	[4,21]	19.12	6.36	[9,29]	2.96	2.54	[0,11]
4	11.84	5.06	[5,21]	18.16	6.46	[4,28]	3.00	2.55	[0, 8]
5	12.12	5.55	[5,24]	17.68	7.10	[5,28]	3.16	2.95	[0,10]
Critical Conditional Difficulty--Level 2									
1	12.92	5.87	[5,26]	17.68	6.54	[5,28]	2.40	2.02	[0, 6]
2	11.60	4.59	[5,20]	18.72	5.86	[9,28]	2.68	2.62	[0, 8]
3	11.96	5.02	[5,25]	18.24	6.11	[3,28]	2.80	2.26	[0, 6]
4	11.92	4.95	[5,21]	18.32	5.89	[9,28]	2.76	2.25	[0, 7]
5	13.60	5.27	[6,26]	17.52	5.70	[7,27]	1.92	1.81	[0, 7]
Critical Conditional Difficulty--Level 3									
1	11.52	4.20	[6,19]	20.16	5.38	[8,27]	1.32	1.85	[0, 7]
2	11.84	4.97	[6,25]	20.00	5.66	[7,27]	1.16	1.41	[0, 5]
3	13.36	4.22	[6,23]	17.64	4.79	[7,27]	2.04	1.48	[0, 5]
4	14.24	4.35	[6,23]	15.80	6.10	[5,27]	3.00	2.86	[0,10]
5	14.88	5.19	[6,27]	15.72	6.10	[4,27]	2.36	2.65	[0,12]
Critical Conditional Difficulty--Level 4									
1	15.92	4.10	[11,25]	14.76	4.95	[5,22]	2.32	2.26	[0, 8]
2	15.40	2.77	[11,21]	15.48	4.00	[8,22]	2.16	2.44	[0,10]
3	14.96	4.60	[9,25]	15.32	5.19	[7,22]	2.68	2.34	[0, 7]
4	14.64	3.92	[9,24]	16.28	5.35	[5,25]	2.12	2.29	[0, 9]
5	16.00	4.41	[10,28]	14.68	4.97	[2,22]	2.36	2.30	[0, 8]

The results of the three-factor analyses of variance for total actual loss with respect to Evaluation Matrices A, B, and C are shown in Tables 9, 10 and 11, respectively. The levels of the variables were analyzed for each significant effect, using the Newman-Keuls test. With respect to evaluation loss matrix A, a significant conditional difficulty level effect was determined ($F=2.11$, $df=6/960$, $p<.05$). As indicated by Table 12, within each model, greater actual loss was obtained for Level 1 than for at least two of the higher levels. Although a significant conditional difficulty level effect with respect to Evaluation Matrix B was found ($F=19.25$, $df=6/960$, $p<.01$), as shown in Table 10, the conditional difficulty level producing significantly less actual loss was Level 1. Table 13 shows that total actual loss increased as the conditional difficulty level increased, for both models, relative to Evaluation Matrix B.

A model effect was determined for Evaluation Loss Matrix B ($F=14.19$, $df=1/960$, $p<.01$). The mean total actual loss for Model 1 was 254.39, whereas for Model 2 the mean was 271.87. Table 14 indicates that although total actual loss was not significantly increased for any two successive levels of the number of items presented prior to an attempt at prediction, generally, as the number of items was increased, the loss increased.

TABLE 9

Analysis of Variance for Total Actual Loss
Related to Evaluation Loss Matrix A

Source	df	MS	F
Model (V_j)	1	3463	1.22
Conditional Difficulty Level ($C_{k(j)}$)	6	5987	2.11*
Items (I_m)	4	5238	1.85
VI_{jm}	4	4191	1.48
$CI_{km(j)}$	24	2317	<1
Error ($\epsilon_{p(jkm)}$)	960	2831	

* $p < .05$.

TABLE 10
 Analysis of Variance for Total Actual Loss
 Related to Evaluation Loss Matrix B

Source	df	MS	F
Model (V_j)	1	74632	14.19**
Conditional Difficulty Level ($C_{k(j)}$)	6	101270	19.25**
Items (I_m)	4	13468	2.56*
VI_{jm}	4	4207	<1
$CI_{km(j)}$	24	886	<1
Error ($\epsilon_{p(jkm)}$)	960	5261	

* $p < .05$.

** $p < .01$.

TABLE 11

Analysis of Variance for Total Actual Loss
Related to Evaluation Loss Matrix C

Source	df	MS	F
Model (V_j)	1	6.89	<1
Conditional Difficulty Level ($C_{k(j)}$)	6	175.31	2.07
Items (I_m)	4	173.54	2.05
VI_{jm}	4	111.86	1.32
$CI_{km(j)}$	24	73.93	<1
Error ($\epsilon_{p(jkm)}$)	960	84.50	

TABLE 12

Results of the Newman-Keuls Test for Total Loss for
Critical Conditional Difficulty Levels Within Model
1 and Model 2 Relative to Evaluation Loss Matrix A

Item	Level			
	1	2	3	4
Model 1				
Means	141.73	123.54	132.36	129.11
Non-significantly different means ^a		+	+	+
Model 2				
Means	131.22	126.46	120.06	134.10
Non-significantly different means ^a	+	+	+	+

Note.-- $\alpha = .05$.

^aMeans are non-significantly different if +'s in any one row correspond to those means.

TABLE 13

Results of the Newman-Keuls Test for Total Loss for
Critical Conditional Difficulty Levels Within Model
1 and Model 2 Relative to Evaluation Loss Matrix B

Item	Level			
	1	2	3	4
Model 1				
Means	223.65	236.06	270.16	288.51
Non-significantly different means	(none)			
Model 2				
Means	245.22	261.18	272.58	308.50
Non-significantly different means	(none)			

Note.-- $\alpha=.05$.

TABLE 14

Results of the Newman-Keuls Test for Total
Loss for Number of Items Presented Prior to
Prediction Relative to Evaluation Loss Matrix B

Item	Number of Items				
	1	2	3	4	5
Means	255.35	258.31	258.54	270.02	273.95
Non-significantly different means ^a	+	+	+	+	+

Note.-- $\alpha=.05$.

^aMeans are non-significantly different if +'s in any one row correspond to those means.

As a means of determining the conditional difficulty level and the number of items that produced the smallest actual loss within each model and with respect to each evaluation matrix, regression analysis was used. The analyses of variance indicated that only with respect to Evaluation Matrix B are both the number of items and the conditional difficulty level effects significant. Also, for Evaluation Matrix B, the item by conditional difficulty level interaction effect was not significant. Therefore a linear regression model, as given by Formula 17, was assumed. For Model 1 the equation obtained was

$$\hat{Y} = 187.84 + 22.87C + 3.19I \quad [21]$$

and for Model 2,

$$\hat{Y} = 201.93 + 20.12C + 6.54I. \quad [22]$$

In Formula 22 both coefficients were not equal to zero ($\alpha = .05$) and the assumed linear model did not appear to be inadequate (Draper & Smith, 1966). The linear model given in Formula 22 did not appear to be inadequate, but one of the coefficients was not significantly different from zero. The standard error of the regression coefficient B_2 was 2.10. The 95 percent confidence interval of the coefficient was $[-0.93, 7.31]$. Since it was possible that the coefficient might be either negative or positive, a conclusion as to the best point estimate, that is, the values for C and I that would minimize total actual loss, was not possible. For this case it was decided that the model given by Formula 18 would then be the assumed model.

For the last case discussed and the remaining four cases, an iterative technique presented by Box and Tidwell (1962) was applied as many times as feasible. In some cases convergence of each of the transformation values was evident, but in others convergence did not seem to occur. In some cases, due to the size of the numbers produced by the transformations, program BMD03R could not be executed again without incurring a computer system error diagnostic, and hence the iterative process was terminated even though convergence was not necessarily evident. Table 15 gives the number of iterations of the transformation process employed, the final transformation values, u and v , and the values of the B_i 's for each of the five cases considered. Table 16 indicates the point estimates for each of the item by conditional difficulty level, obtained from each of the resulting regression equations. In Table 16, for each evaluation matrix within a model, the minimal predicted actual loss has been indicated. The conditional difficulty by item level producing the minimal predicted loss for each evaluation matrix was to be identified for use in part 2 of the study. Only for Evaluation Matrix C within Model 2 was the minimal loss value not unique. Because of the extensive amount of time and expense involved in running the computer simulations, conditional difficulty level 3 and item level 1 were selected for Matrix C within Model 2 for use in the second part of the study. By making this selection, one set of simulations could be run and evaluated with respect to both Evaluation Matrices

TABLE 15
Results of Regression After Transformation
of the Independent Variables

Evaluation loss Matrix	Number of Iterations	Transformation Values		Constant (B_0)	Regression Coefficients				
		u	v		B_1	B_2	B_{11}	B_{22}	
									B_{12}
Model 1									
A	2	-10.098	0.946	134.27	-7925.81	-4.94	7937.25	1.10	-0.18
B	3	-1.478	7.688	326.78	-351.34	0.00034	239.38	0.00000	0.00012
C	1	-3.475	1.369	20.646	-30.474	-0.150	29.875	0.024	0.087
Model 2									
A	7	3.968	-2.742	137.49	-0.22	-91.22	0.00092	76.14	0.037
C	1	2.931	15.758	19.026	-0.105	0.00000	0.0023	0.00000	0.00000

Note.--Assumed Model-- $Y = B_0 + B_1 C^u + B_2 I^v + B_{11} (C^u)^2 + B_{22} (I^v)^2 + B_{12} C^u I^v + \epsilon$.

TABLE 16

Point Estimates of Total Actual Loss for Conditional
Difficulty by Item Levels Within Each Model

Condi- tional Diffi- culty Level	Item Level	Model 1			Model 2		
		Matrix			Matrix		
		A ^a	B ^a	C ^a	A ^a	B ^b	C ^a
1	1	141.68	214.82 ^C	20.01	122.23	228.60 ^C	18.92
1	2	139.91	214.91	20.05	125.34	235.14	18.92
1	3	139.99	216.96	20.25	132.97	241.68	18.92
1	4	141.80	234.36	20.71	135.27	248.23	18.92
1	5	145.26	323.48	21.46	136.17	254.77	18.92
2	1	123.20	231.48	18.03	119.79	248.72	18.36
2	2	121.59 ^C	231.56	17.94 ^C	122.43	255.27	18.36
2	3	121.83	233.26	18.00	130.01	261.81	18.36
2	4	123.79	247.76	18.28	132.29	268.35	18.36
2	5	127.41	321.98	18.84	133.19	274.90	18.36
3	1	130.30	266.79	19.87	113.77 ^C	268.85	17.84 ^C
3	2	128.69	266.86	19.77	114.46	275.39	17.84
3	3	128.93	268.48	19.81	121.80	281.93	17.84
3	4	130.90	282.24	20.08	124.03	288.48	17.84
3	5	134.51	352.70	20.63	124.91	295.02	17.84
4	1	130.41	285.45	20.28	132.83	288.97	20.72
4	2	128.80	285.52	20.18	128.32	295.52	20.72
4	3	129.05	287.10	20.22	135.05	302.06	20.72
4	4	131.01	300.56	20.48	137.11	308.60	20.72
4	5	134.63	369.42	21.03	137.93	315.15	20.72

^aRegression model-- $Y = B_0 + B_1C^U + B_2I^V + B_{11}(C^U)^2 + B_{22}(I^V)^2 + B_{12}C^UI^V + \epsilon$.

^bRegression model-- $Y = B_0 + B_1C + B_2I + \epsilon$.

^cMinimal point estimate for the matrix within the model.

A and C. A similar situation occurred within Model 1. One set of simulations was run for conditional difficulty level 2 and item level 2, but the results were evaluated with respect to both Matrix A and Matrix C.

Part 2 of the Study

Table 17 indicates the descriptive statistics for total actual loss incurred for each evaluation matrix within each model and for each size of data base for which a computer simulation was run.

Table 18 shows the correlation between total actual loss incurred with a 2000-data base and a data base of size e for each evaluation matrix within each model. The table also indicates the computed t -values, the probabilities of obtaining the t -values, and the conclusions with respect to the null hypothesis

$$H_0: \bar{L}_e - \bar{L}_{2000} = 0$$

and the alternate hypothesis

$$H_1: \bar{L}_e - \bar{L}_{2000} > 0,$$

where $\bar{L}_e$ and $\bar{L}_{2000}$ are the mean actual losses for simulations run with data bases of size e and 2000, respectively. As indicated in Table 18, all null hypotheses tested were rejected except for two cases, both of which involved Matrix B.

The final purpose of the study was to determine for each of the evaluation loss matrices within each model whether a tailored testing procedure employing a data base of fewer than 1000 sets of responses would be more efficient than presenting all the items.

TABLE 17

Descriptive Statistics of Total Actual Loss
Involving Data Bases of Different Sizes

Loss Matrix	Size of Data Base							
	100		200		1000		2000	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Model 1								
A	144.37	68.73	138.39	62.89	134.73	68.35	126.04	56.66
B	219.55	79.92	207.68	59.87	---	---	204.22	57.57
C	20.98	11.30	20.51	10.59	19.89	11.46	18.59	9.75
Model 2								
A	153.52	72.63	146.51	66.16	134.04	45.33	128.72	47.97
B	276.34	103.54	---	---	---	---	270.92	92.91
C	23.72	12.27	22.62	11.19	20.94	8.04	19.87	8.56

TABLE 18

Comparison of Actual Losses for Simulations Employing Data Bases of Size e with Losses Obtained by Using a Data Base of Size 2000

Data Base Size (e)	Correlation	t	p	Conclusion ^{a,b}
Matrix A--Within Model 1				
100	.76	4.08	< .01	Reject H_0
200	.81	3.28	< .01	Reject H_0
1000	.77	1.99	< .05	Reject H_0
Matrix B--Within Model 1				
100	.64	2.48	< .01	Reject H_0
200	.70	0.75	> .05	Fail to reject H_0
Matrix C--Within Model 1				
100	.79	3.38	< .01	Reject H_0
200	.85	3.35	< .01	Reject H_0
1000	.79	1.85	< .05	Reject H_0
Matrix A--Within Model 2				
100	.70	4.75	< .01	Reject H_0
200	.65	3.52	< .01	Reject H_0
1000	.82	1.88	< .05	Reject H_0
Matrix B--Within Model 2				
100	.68	0.68	> .05	Fail to reject H_0
Matrix C--Within Model 2				
100	.76	4.79	< .01	Reject H_0
200	.73	3.56	< .01	Reject H_0
1000	.85	2.34	< .01	Reject H_0

^aNull hypothesis (H_0)-- $\bar{L}_e - \bar{L}_{2000} = 0$
 Alternate hypothesis (H_1)-- $\bar{L}_e - \bar{L}_{2000} > 0$.

^bAlpha level of each test = .05.

The smallest data base considered was of size 100. The hypotheses tested were of the form

$$H_0: \bar{L}_{100} - \bar{L} = 0$$

with respect to an alternate hypothesis of the form

$$H_1: \bar{L}_{100} - \bar{L} < 0,$$

where L was the computed minimum loss for asking all the items. It has previously been mentioned that these computed values for L were 165, 495, and 33, for Evaluation Matrices A, B, and C, respectively. These values may be compared with the actual mean losses incurred for data bases of varying sizes by inspecting the means listed in Table 17. Table 19 provides a tabulation of the computed t -value for each statistical test, and the probability of obtaining each t -value. Since the null hypothesis was not rejected for Matrix A within Model 2, the hypothesis

$$H_0: \bar{L}_{200} - 165 = 0$$

was tested at an α -level of .01 with respect to the alternate hypothesis

$$H_1: \bar{L}_{200} - 165 < 0.$$

The computed t -value was -2.79 ($p < .005$). Hence the null hypothesis was rejected.

TABLE 19

Results of One-Tailed t-tests to Compare Total
Actual Loss Obtained for Data Bases of Size
100 and Loss Incurred for Presenting All Items

Evaluation Loss Matrix	t	p
Model 1		
A	- 4.21	< .001
B	- 34.06	< .001
C	- 10.58	< .001
Model 2		
A	- 1.58	> .05
B	- 21.12	< .001
C	- 7.58	< .001

Note.--Null hypothesis-- $\bar{L}_{100} - L = 0$, alternate hypothesis-- $\bar{L}_{100} - L < 0$, where L equals 165, 495, or 33, for matrices A, B, and C respectively.

DISCUSSION

The dependent variables used for every statistical test were total actual losses, which were computed on the basis of the values in the three evaluation matrices. The matrices are shown in Figures 8, 9, and 10. The corresponding values in the matrices differ. Evaluation of a tailored testing procedure using differing evaluation matrices might produce different conclusions concerning the efficiency of the testing procedure. Such was the case in this study.

Evaluation Matrices A and C, shown in Figures 2 and 4, respectively, have different corresponding values, but the values specified in each matrix indicate similar comparative loss patterns. For both matrices the greatest expected loss is specified for the action, predicting an examinee's response to be correct when the item response, if obtained, would have been wrong. For the other possible error in prediction, that is, predicting an examinee's response to be wrong when the item response would have been correct, the expected loss specified in each matrix was two-thirds the expected loss for the former prediction error and greater than any of the other specified loss values within their respective matrices. The smallest expected loss values in each matrix were assigned to the cases

indicating correct prediction. Both matrices indicated that expected losses for obtaining the examinee's response would be greater than expected losses for correct predictions, but less than expected losses for incorrect predictions.

Evaluation Matrix B, shown in Figure 9, indicates that the greatest losses were assigned to the categories corresponding to obtaining an examinee's response. As in Matrices A and C, Matrix B has the lowest expected loss values for cases involving correct prediction. Although in Matrix B the loss values for incorrect prediction are greater than the values for correct prediction, they are lower than losses expected for obtaining an examinee's response.

With respect to all three evaluation matrices, the most efficient tailored testing procedure would be one in which all the item responses would be correctly predicted. With respect to Matrices A and C, a more efficient testing procedure would be one in which all item responses had been obtained rather than one in which all the responses had been incorrectly predicted; whereas, with respect to Matrix B, the opposite is true. The author does not concur with the relative losses indicated in Matrix B. The reader may also fail to concur with the values specified in any one of the matrices. It was not a purpose of the study to determine what the values should be or which of the three evaluation matrices is most appropriate to use, but rather it was intended that the ramifications of employing different levels of the variables considered in the study were to be considered with respect to the differing evaluative criteria.

Which of the Two Tailored Testing Models is More Efficient?

With respect to Evaluation Matrices A and C, both models provided equally efficient testing procedures. Although time was not considered as a variable in the study, the variable may be very important in relation to the testing procedures. It was evident to the investigator that the amount of computer processing time needed to run the computer simulations for Model 2 was much more than that needed for Model 1. Prior to selecting one of the models for implementation, it would appear to be appropriate to consider processing time, cost, and delay of response by the computer to the student for each of the models.

With respect to Evaluation Matrix B, Model 1 was more efficient than Model 2. This result was probably due to the necessity of obtaining more responses to items in Model 2 than in Model 1. Comparing the number of items presented for each item level within each critical conditional difficulty level, the values for Model 1, displayed in Table 7, generally appear to be higher than the corresponding values for Model 2, shown in Table 8. As previously discussed, the expected losses for obtaining responses, specified in Matrix B, were greater than for any other conditions. Hence, it would be expected that if the models were not equally as efficient, Model 1 would be superior.

How Many Items Should be Presented Before Attempting Prediction?

With respect to Evaluation Matrices A and C, presenting one or as many as five items prior to predicting item responses

produced equally efficient testing procedures. Although there was no difference in efficiency, as measured in terms of total loss, for the different number of items presented prior to prediction, the presentation of more than one item prior to prediction might increase the time needed by an examinee to complete a tailored testing session. If a purpose of the tailored testing procedure is to reduce the amount of time needed to test the student, consideration should be given to the effects of the number of items presented initially on total time needed for testing.

Although with respect to Evaluation Matrices A and C no difference in total loss resulted from increasing the number of items presented prior to prediction, with respect to Matrix B, the total loss increased as the number of items initially presented increased. Hence, with respect to Matrix B, the most efficient tailored testing procedures required the presentation of one item prior to prediction.

What Critical Conditional Difficulty Values Should be Selected?

Model 1. With respect to Matrix A, the lowest critical conditional difficulty level (Level 1--critical values of .2 and .8) produced the greatest total loss. The other critical conditional difficulty levels produced equivalent losses.

With respect to Matrix B, total loss increased as the critical conditional difficulty level increased. Hence, the lowest

critical conditional difficulty level produced the most efficient testing procedures.

Unlike the results obtained for Matrices A and B, no critical conditional difficulty level provided smaller total losses than any other level for Matrix C. Although Matrix C was similar to Matrix A in comparative loss patterns, there were greater differences in the magnitudes of pairs of loss values in Matrix A than for most corresponding pairs of loss values in Matrix C.

Model 2. With respect to Matrix A, Conditional Difficulty Level 3 produced minimal total losses. As with Model 1, for Matrix B, total losses increased as the conditional difficulty level increased. Total losses were minimized when $V_1 = .20$ and $V_2 = .80$. For Matrix C, all critical conditional difficulty levels produced equivalent total losses.

Does a Data Base Consisting of Fewer than 1000 Sets of Item Responses Provide as Efficient a Procedure as a Data Base Consisting of 2000 Sets of Responses?

Only with respect to Evaluation Matrix B did data bases consisting of fewer than 1000 sets of item responses provide as efficient a procedure as a data base consisting of 2000 sets of responses. For Model 1, a data base of 200 sets of responses proved to be as efficient as the 2000-data base; for Model 2, a data base as small as 100 sets of responses proved to be as efficient.

Are the Tailored Testing Procedures, Using Data Bases Consisting of Fewer than 1000 Sets of Item Responses, More Efficient than Presenting All the Items to an Examinee?

For both versions of the model and with respect to all evaluation matrices, data bases of sizes less than 1000 sets of item responses provided more efficient testing procedures than presenting all the items of the test to the examinee. Data bases of size 100 provided for superior efficiency of the tailored testing procedures in every case except for Evaluation Matrix A within Model 2. For this one case, a data base of 200 sets of responses was necessary to provide for a tailored testing procedure that was more efficient than presenting all the items to an examinee.

Implications of the Study

The purpose of the study was to determine which of two versions of a tailored testing procedure and what levels of the variables in the testing models produced the most efficient procedure, and what size data base produced a more efficient testing procedure than presenting all the items to an examinee. The conclusions reached differed according to the criteria used to evaluate the testing procedures. Even for similar, although not exactly the same, criteria, conclusions reached were not always the same. Also, the conclusions reached were with respect to one set of test items. Other individuals, using a different set of items, would not necessarily agree to the values used in any of the evaluation matrices employed in the study. Hence, it would appear that simulations similar to those conducted in this

study or studies using human subjects would be appropriate, in order for the decision makers to determine what levels of the variables produce the most efficient procedure.

Although the use of different evaluation criteria resulted in different conclusions, for all three evaluation matrices it was shown that both versions of the tailored testing model could be used with data bases of sizes 100 or 200 to produce testing procedures that were more efficient than presenting all the items to an examinee. Although larger data bases may result in smaller total actual loss, the use of data bases of sizes 100 or 200 may provide a means for tailored testing that is more effective than requiring examinees to answer all the items of a test.

Eliminating the Decision Loss Functions

Formulas 11 and 12 indicate that expected loss for prediction is a function of the conditional difficulty of an item and the critical loss value, which is dependent upon the critical conditional difficulty selected. The expected loss for obtaining a response, as indicated in Formula 13, is also dependent upon the conditional difficulty of an item and the size of the data base upon which the conditional difficulty has been computed. With such dependence of expected loss values on conditional difficulty, one wonders if the decision loss functions might be replaced by decision making based solely upon the conditional difficulty of an item. Even

though the decisions are affected by the size of the data base, the range in adjusted critical values may not be sufficiently large to affect the results of the tailored testing procedures. An argument to eliminate the decision loss functions is indeed tenable. But one should also note that the nearly complete dependence of the expected loss values on conditional difficulty is a result of the assumed loss functions--two linear functions and a quadratic function. The testing models do not require the use of these particular functions. Further research may reveal other loss distributions that provide greater efficiency. Hence, it appears that there are two nearly opposite directions for further investigation of the models. The first is to replace the decision loss matrix and corresponding functions with a decision making strategy based directly on conditional difficulty values. The second direction is the investigation of other loss distributions.

Applicability of the Models to Instructional Testing

As previously indicated, the primary intent of the development of the proposed tailored testing models has been for the purpose of providing an efficient computer-based means of testing in instructional situations. Kalisch (1974) indicated the possible applicability of his testing model to determination of mastery of objectives, but no method of assessing mastery was advanced. It appears that with the strong emphasis on objective-referenced testing and mastery learning in present day education, it is

necessary to be able to express the results of the tailored testing procedures in terms of mastery of objectives. Two methods are suggested for possible future research. The first requires no change to the existing models. After employing either testing procedure, each of the responses, either obtained or predicted, for the items corresponding to an objective would be analyzed to determine if the minimum percentage of items would have been answered correctly. The second method would involve prediction of mastery of each objective after each obtained response to an item. With this latter method, conditional probabilities of the mastery of objectives, given the examinee's item responses, would be computed.

It is not the intent of the author to discuss in this paper the numerous practical considerations that must be made prior to implementing a tailored testing model in an instructional situation. The results of the study indicate that the models are feasible for predicting the correctness of item responses, employing relatively small data bases, sizes of 100 or 200. On the basis of the results of this study and Kalisch's (1974) previous study, additional investigations into the refinement and extension of the tailored testing models would appear to be worthwhile.

REFERENCES

- Box, G. E. P., & Tidwell, P. W. Transformation of the independent variables. Technometrics, 1962, 4, 531-550.
- Cohen, J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1969.
- Cronbach, L., & Gleser, G. Psychological tests and personnel decisions. Urbana, Ill.: University of Illinois Press, 1965.
- Dixon, W. J. (Ed.) Biomedical computer programs. Berkley: University of California Press, 1971.
- Hays, W., & Winkler, R. Statistics: Probability, inference, and decision. Vol. 1. New York: Holt, Rinehart & Winston, 1970.
- Kalisch, S. J. A tailored testing model employing the beta distribution and conditional difficulties. Journal of Computer-Based Instruction, 1974, 1, 22-28.
- Veldman, D. Fortran programming for the behavioral sciences. New York: Holt, Rinehart & Winston, 1967.